A Consistent Estimator of the Evolutionary Rate

Krzysztof Bartoszek and Serik Sagitov

bioRxiv first posted online September 2, 2014
Access the most recent version at doi: http://dx.doi.org/10.1101/008714

A Consistent Estimator of the Evolutionary Rate

Krzysztof Bartoszek and Serik Sagitov

September 2, 2014

Abstract

We consider a branching particle system where particles reproduce according to the pure birth Yule process with the birth rate λ, conditioned on the observed number of particles to be equal n. Particles are assumed to move independently on the real line according to the Brownian motion with the local variance σ^2. In this paper we treat n particles as a sample of related species. The spatial Brownian motion of a particle describes the development of a trait value of interest (e.g. log–body–size). We propose an unbiased estimator R_n^2 of the evolutionary rate $\rho^2 = \sigma^2/\lambda$. The estimator R_n^2 is proportional to the sample variance S_n^2 computed from n trait values. We find an approximate formula for the standard error of R_n^2 based on a neat asymptotic relation for the variance of S_n^2.

(Keywords: Branching Brownian motion, conditioned branching process, tree–free phylogenetic comparative method, quantitative trait evolution, Yule process)

1 Introduction

Biodiversity within a group of n related species could be quantified by comparing suitable trait values. For some key trait values like log body size, researchers apply the Brownian motion model proposed by Felsenstein [1985]. It is assumed that the current trait values $(X_1^{(n)},\ldots,X_n^{(n)})$ have evolved from the common ancestral state X_0 as a branching Brownian motion with the local variance σ^2. Given a phylogenetic tree describing the ancestral history of the group of species the Brownian trajectories of the trait values for sister species are assumed to evolve independently after the ancestor species splits in two daughter species. The resulting phylogenetic sample $(X_1^{(n)},\ldots,X_n^{(n)})$ consists of identically distributed normal random variables with a dependence structure caused by the underlying phylogenetic signal.

A mathematically appealing and biologically motivated version of the phylogenetic sample model assumes that the phylogenetic tree behind the normally distributed trait values $(X_1^{(n)},\ldots,X_n^{(n)})$ is unknown. As a natural first choice to model the unknown species tree, we use the Yule process with birth rate λ [see Yule, 1924]. Since the phylogenetic sample size is given, n, the Yule process should be conditioned on having n tips: such conditioned branching processes have received significant attention in recent years, due to e.g. Aldous and Popovic [2005], Gernhard [2008], Mooers et al. [2012], Stadler [2009, 2011], Stadler and Steel [2012]. This ”tree-free” approach for comparative phylogenetics was previously addressed by Sagitov and Bartoszek [2012] and Crawford and Suchard [2013], [much earlier Edwards, 1970, used a related branching Brownian process as a population genetics model].

In our work we show that a properly scaled sample variance is an unbiased and consistent estimator of the compound parameter $\rho^2 = \sigma^2/\lambda$ which we call the evolutionary rate of the trait value in question. Our main mathematical result, Theorem 2.1, gives an asymptotical expression for the variance of the phylogenetic sample variance. This result leads to a simple asymptotic formula for the estimated standard error of our estimator. Our result is in agreement with the work of Crawford and Suchard [2013] whose simulations indicate that their approximate maximum likelihood procedure yields an unbiased consistent estimator of σ^2. This is illustrated using the example of the Carnivora order studied previously by Crawford and Suchard [2013].

The phenotype modelled by a Brownian motion is usually interpreted as the case of neutral evolution with random oscillations around the ancestral state. This model was later developed into an adaptive evolutionary model based on the Ornstein–Uhlenbeck process by Felsenstein [1988], Hansen [1997], Butler and King [2004], Hansen et al. [2008], Bartoszek et al. [2012]. The tree-free setting using the Ornstein–Uhlenbeck process was addressed by Bartoszek and Sagitov [2012] where for the Yule–Ornstein–Uhlenbeck model, some phylogenetic confidence intervals for the optimal trait value were obtained via three limit theorems for the phylogenetic sample mean. Furthermore, it was shown that the phylogenetic sample variance is an unbiased consistent estimator of the stationary variance of the process.

At the end of their discussion Crawford and Suchard [2013] write that as the the tree of life is refined interest in “tree–free” estimation methods may diminish. They however indicate that “tree–free” estimates may be useful to calculate starting points for simulation analysis. We certainly agree with the second statement but believe that development of “tree–free” methods should proceed alongside that of “tree–based” ones.

One of the most useful features of the tree–free comparative models is that they offer a natural method of tree growth allowing for study of theoretical properties of phylogenetic models as demonstrated in this work [and also Sagitov and Bartoszek, 2012, Bartoszek and Sagitov, 2012, Bartoszek, 2014, Crawford and Suchard, 2013]. Another alternative to studying properties of these estimators is the tree growth model proposed by Ané [2008], Ho and Ané [2013], Ané et al. [2014]. In this setup the total height of the tree is kept fixed and new tips are added to randomly chosen branches. These two approaches seem to be in agreement, at least up to the second moments, since e.g. they agree on the lack of consistency of estimating X_0. In Sagitov and Bartoszek [2012] we showed that under the Yule Brownian motion model $\mathrm{Var}\left[\overline{X}_n\right] \to 2\sigma^2$.

In a practical situation “tree–free” methods can be used for a number of purposes. Firstly as pointed out by Crawford and Suchard [2013] they can be useful for calculating starting points for further numerical estimation procedures or defining prior distributions in a Bayesian setting. Secondly they have to be used in a situation where the tree is actually unknown e.g. when we are studying fossil data or trying to make predictive statements about future phenotypes, e.g. development of viruses. Thirdly they can be used for various sanity checks. If they contradict “tree–based” results this could indicate that the numerical method fell into a local maximum.

The paper has the following structure. Section 2 presents the model, the main results and an application. Section 3 states two lemmata and a proposition directly yielding the assertion of Theorem 2.1. Proposition 3.1 deals with the covariances between coalescent times for randomly chosen pairs of tips from a random Yule n-tree. The properties of the coalescent of a single random pair were studied previously by e.g. Steel and McKenzie [2001] and Sagitov and Bartoszek [2012]. In Section 4 we state two lemmata needed for the proof of Proposition 3.1. Section 5 contains two further lemmata and the proof of Proposition 3.1. In Section 6, 7, and 8 we prove the lemmata from Sections 3, 4, and 5. Appendix A contains some useful results concerning harmonic numbers of the first and second order.

2 The main results

The basic evolutionary model considered in this paper is characterized by four parameters (λ,n,X_0,σ^2) and consists of two stochastic components: a random phylogenetic tree defined by parameters (λ,n) and a trait evolution process along a lineage defined by parameters (X_0,σ^2). The first component, species tree connecting n extant species, is modelled by the pure birth Yule process [Yule, 1924] with the birth (speciation) rate λ and conditioned on having n tips [Gernhard, 2008]. For the second component we adapt the approach by assuming that for a given $i=1,\ldots,n$, the current trait value $X_i^{(n)}$ has evolved from the ancestral state X_0 according to the Brownian motion with the local variance σ^2.

Treating the collection of the current trait values $(X_1^{(n)},\ldots,X_n^{(n)})$ generated by such a process as a sample of identically distributed, but dependent, observations, we are interested in the properties of the basic summary statistics

$$\overline{X}_n=\frac{X_1^{(n)}+\ldots+X_n^{(n)}}{n},\quad S_n^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i^{(n)}-\overline{X}_n)^2,$$

the sample mean and sample variance.

According to [Sagitov and Bartoszek, 2012] we have

$$\mathrm{E}\left[S_n^2\right]=\left(\frac{n+1}{n-1}H_n-2\frac{n}{n-1}\right)\frac{\sigma^2}{\lambda},$$

see Fig 1, left panel (all simulations are produced using the TreeSim [Stadler, 2009, 2011] and mvSLOUCH [Bartoszek et al., 2012] R packages). It follows that the normalized sample variance

$$R_n^2=\left(\frac{n+1}{n-1}H_n-2\frac{n}{n-1}\right)^{-1}S_n^2 \tag{1}$$

gives an unbiased estimator of the compound parameter $\rho^2:=\frac{\sigma^2}{\lambda}$ for the Yule–Brownian–Motion model, see Fig 2. In the comparative phylogenetics framework the ratio ρ^2 can be called the *evolutionary rate* as it measures the speed of change in the trait value when the time scale is such that we expect one speciation event per unit of time and per species. The next theorem is the main asymptotic result of this paper, illustrated by Fig 1, right panel.

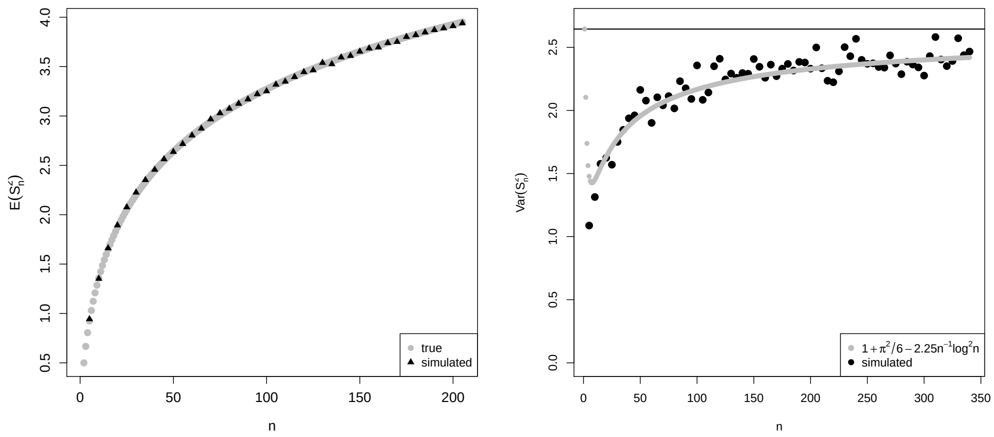

Figure 1: Left: True and simulated values of E $\left[S_n^2\right]$, right: simulated values of Var $\left[S_n^2\right]$ with limit equalling $\pi^2/6+1$. Each point comes from 10000 simulated Yule trees and Brownian motions on top of them. Parameters used in simulations are $\lambda=1$, $X_0=0$ and $\sigma^2=1$. The grey line on the right panel fits a curve based on the convergence rate $O(n^{-1}\log n^2)$.

Theorem 2.1 *Consider the sample variance S_n^2 for the Yule–Brownian–Motion model with parameters $(\lambda, n, X_0, \sigma^2)$. Its variance satisfies the following asymptotic relation*

$$\mathrm{Var}\left[S_n^2/\rho^2\right] = 1+\frac{\pi^2}{6}+O(n^{-1}\log^2 n), \quad n\to\infty.$$

In terms of our estimator (1), Theorem 2.1 yields

$$\mathrm{Var}\left[R_n^2/\rho^2\right] = \frac{1+\frac{\pi^2}{6}}{(\log n+\gamma-2)^2}+O(n^{-1}),$$

where $\gamma=0.577$ is the Euler constant, implying that R_n^2 is a consistent estimator of the evolutionary rate ρ^2. It follows that for large n, the standard error (estimated standard deviation) of the unbiased estimator R_n^2 can be approximated by

$$\mathrm{SE}(R_n^2)\approx\sqrt{1+\frac{\pi^2}{6}}\cdot\frac{R_n^2}{\log n+\gamma-2}\approx\frac{1.626}{\log n-1.423}\cdot R_n^2. \qquad (2)$$

The estimator of Eq. (1) should be compared to the approximate maximum–likelihood estimator for the local variance σ^2 recently proposed by Crawford and Suchard [2013] in the same framework of the Yule–Brownian–Motion model. The main difference between two approaches is that in Crawford and Suchard [2013] it is assumed that one knows both the number of tips and the total height of the otherwise unknown species tree. The Crawford-Suchard estimator is based on a closed form of the distribution of phylogenetic diversity – the sum of branch lengths connecting the species in a clade.

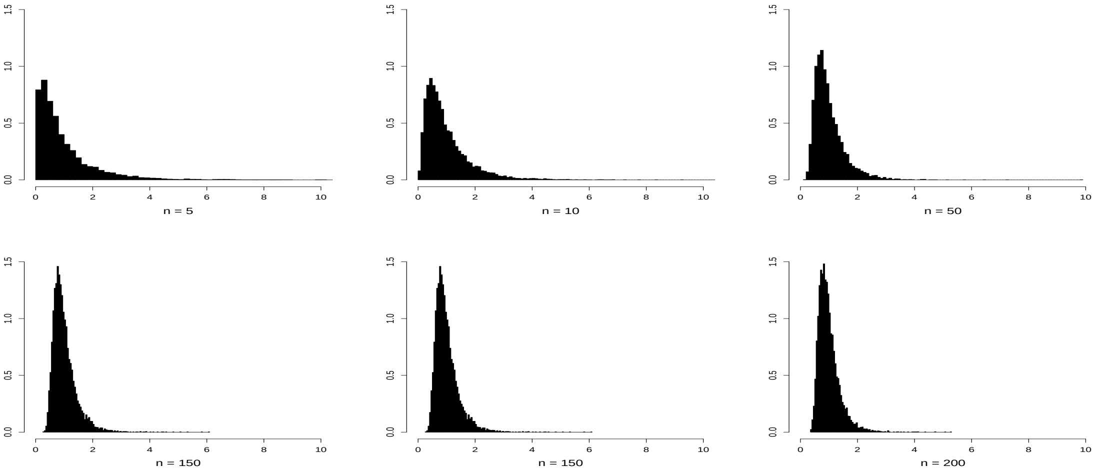

Figure 2: Histograms of R_n^2 for left to right top $n = 5, 10, 50$ and bottom $n = 100, 150, 200$. Parameters used in simulations are $\lambda = 1$, $X_0 = 0$ and $\sigma^2 = 1$.

As an application of their estimator, Crawford and Suchard [2013] study different families of the Carnivora order, estimating σ^2 for each of the 12 clades. The data for the log-body-size disparities was taken from the PanTHERIA database [Jones et al., 2009]. The data summary and the Crawford-Suchard estimates are shown in the left part of Tab. 1. In the right part of Tab. 1 we present our estimates $\hat{\rho}^2$ for the evolutionary rate parameter $\rho^2 = \sigma^2/\lambda$ for each of the 12 families in the Carnivora order. The standard error is computed using (2). We note that the data does not take into account the newly described species *Bassaricyon neblina* from the Procyonidae family [Helgen et al., 2013].

In the next-to-last column we list the ratios demonstrating a surprisingly good agreement between our and Crawford-Suchard estimates. The ratio is taken between two products: $\hat{\rho}^2 u_n$ on one hand, and $\hat{\sigma}^2 t_n$ on the other. Here $u_n = \mathrm{E}[U_n]$ is the expected age of the conditioned standard Yule process with $\lambda = 1$, while t_n is the clade age assumed to be known in the Crawford-Suchard framework. Both $\hat{\rho}^2 u_n$ and $\hat{\sigma}^2 t_n$ estimate the same quantity – the variance in the trait values for the evolution of the corresponding clade. Therefore, one should expect these ratios to be close to one. And indeed, the 12 ratios have mean 0.97 and standard deviation 0.20.

Our estimator and its standard error are computed by simple formulae given above. A major weakness of our estimator is relatively big standard error for realistic richness values, see the 7th column in Tab. 1. This can be explained by the fact that we do not use an additional information about the species tree, like the height of the tree used in the Crawford-Suchard estimator.

This close agreement is obtained despite a number of features that complicates the comparison between two methods. Our approach in its current form does not allow to take into account the fact that some trait values are missing. We calculated $\hat{\rho}^2$ for the trait disparity as if it was computed using all n trait values. Moreover, it is not be clear how to take into account the measurement variance. As shown by Hansen and Bartoszek [2012] even with a known tree, the measurement

Family	n	t_n	Disparity	$\hat{\sigma}^2$ (SE)	u_n	$\hat{\rho}^2$ (SE)	$\frac{\hat{\rho}^2 u_n}{\hat{\sigma}^2 t_n}$	$\frac{\hat{\rho}^2}{\hat{\sigma}^2/\hat{\lambda}}$
Felidae	40 (7)	33.3	1.588	.080 (.009)	4.279	.649 (.466)	1.042	0.560
Viverridae	35 (6)	37.4	0.662	.029 (.004)	4.147	.284 (.217)	1.086	0.676
Herpestidae	33 (4)	25.5	0.482	.030 (.003)	4.089	.211 (.166)	1.128	0.485
Eupleridae	8 (0)	25.5	0.916	.079 (.010)	2.718	.758 (1.72)	1.023	0.662
Hyaenidae	4 (0)	32.2	0.805	.122 (.005)	2.083	.999 (19.5)	0.530	0.565
Canidae	35 (3)	48.9	0.678	.030 (.004)	4.147	.290 (.221)	0.825	0.667
Ursidae	8 (0)	42.6	0.303	.024 (.002)	2.718	.251 (.569)	0.667	0.722
Otariidae	16 (2)	24.5	0.386	.028 (.003)	3.381	.227 (.274)	1.119	0.559
Phocidae	19 (0)	24.5	0.751	.052 (.005)	3.548	.410 (.438)	1.142	0.544
Mephitidae	12 (3)	32.0	0.570	.039 (.005)	3.103	.384 (.588)	0.955	0.679
Mustelidae	59 (10)	27.4	2.263	.126 (.014)	4.663	.811 (.497)	1.095	0.444
Procyonidae	14 (1)	27.4	0.531	.037 (.004)	3.252	.332 (.444)	1.065	0.619

Table 1: Data summary. 2nd column: clade richness (number of missing trait values); 3rd column: the clade age in millions of years; 4th column: $\frac{n-1}{n} \cdot S_n^2$ trait disparity ; 6th column: the expected age $u_n = \mathrm{E}\left[U_n\right]$ of the conditioned standard Yule process with $\lambda = 1$.

error can cause very diverse effects. Therefore we would expect the situation to be even more interesting when we integrate the phylogeny out.

In their work Crawford and Suchard [2013] estimated the overall speciation rate to be $\hat{\lambda} = 0.069$ per million years. The last column of Tab. 1 demonstrates that using this common value for the speciation rate λ produces huge discrepancy between our estimates $\hat{\rho}^2$ for the rates of evolution $\rho^2 = \sigma^2/\lambda$ and the rates of evolution computed using the Crawford-Suchard estimates for σ^2. This observation points out that a fair direct comparison of $\hat{\rho}^2$ and $\hat{\sigma}^2/\hat{\lambda}$ would requires specific estimates of the speciation rate λ for each of the 12 clades.

3 Outline of the proof of Theorem 2.1

We start with a general observation, Lemma 3.1, concerning the sample variance

$$D_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

of n, possibly dependent and not necessarily identically distributed, observations $(Y_1, \ldots, Y_n)$ with sample mean $\overline{Y} = n^{-1}\sum\limits_{i=1}^{n} Y_i$.

Lemma 3.1 *If (W_1, W_2, W_3, W_4) is a random sample without replacement from random values $(Y_1, \ldots, Y_n)$, then*

$$\mathrm{Var}\left[D_n^2\right] = \mathrm{Cov}\left[W_1^2, W_2^2\right] - 2\,\mathrm{Cov}\left[W_1^2, W_2W_3\right] + \mathrm{Cov}\left[W_1W_2, W_3W_4\right] + n^{-1}B_n, \qquad (3)$$

where

$$|B_n| < \mathrm{E}\left[W_1^4\right] + 4\,\mathrm{E}\left[W_1^3 W_2\right] + \mathrm{E}\left[W_1^2 W_2^2\right] + 6\,\mathrm{E}\left[W_1^2 W_2 W_3\right] + 4\,\mathrm{E}\left[W_1 W_2 W_3 W_4\right].$$

Observe that in terms of the sample variance for the scaled trait values

$$Y_i := Y_i^{(n)} = \frac{X_i^{(n)} - X_0}{\sigma/\sqrt{\lambda}}, \quad i = 1, \dots, n, \tag{4}$$

we have $S_n^2 = \frac{\sigma^2 D_n^2}{\lambda}$, and to prove Theorem 2.1 we have to verify that

$$\mathrm{Var}\left[D_n^2\right] = 1 + \frac{\pi^2}{6} + O(n^{-1}\log^2 n). \tag{5}$$

The Yule n-tree underlying the set of scaled values (4) has unit speciation rate. We call it the *standard* Yule n-tree, and denote by $\mathscr{Y}_n$ be the σ–algebra generated by all the information describing this random tree. Under the Brownian motion assumption the trait values (4) are conditionally normal with

$$\mathrm{E}\left[Y_i|\mathscr{Y}_n\right] = 0, \qquad \mathrm{Var}\left[Y_i|\mathscr{Y}_n\right] = U_n,$$

where U_n is the height of the standard Yule n-tree, see Fig. 3. Moreover, see Section 6, we have

$$\mathrm{Cov}\left[Y_i, Y_j|\mathscr{Y}_n\right] = U_n - \tau_{ij}^{(n)}, \tag{6}$$

where $\tau_{ij}^{(n)}$ is the backward time to the most recent common ancestor for a pair of distinct tips (i, j) in the standard Yule n-tree, see Fig. 3. For a quadruplet (i, j, k, l) of tips randomly sampled without replacement out of n tips in the standard Yule n-tree, we denote

$$\tau_1^{(n)} = \tau_{ij}^{(n)}, \quad \tau_2^{(n)} = \tau_{ik}^{(n)}, \quad \tau_3^{(n)} = \tau_{lk}^{(n)}, \quad \tau_4^{(n)} = \tau_{jk}^{(n)}, \quad \tau_5^{(n)} = \tau_{jl}^{(n)}, \quad \tau_6^{(n)} = \tau_{kl}^{(n)}. \tag{7}$$

Lemma 3.2 *Let* (W_1, W_2, W_3, W_4) *be a random sample without replacement of four trait values out of n random values defined by* (4)*. Then in terms of the coalescent times* (7) *we have*

$$\begin{aligned}\mathrm{Cov}\left[W_1^2, W_2^2\right] - 2\,\mathrm{Cov}\left[W_1^2, W_2 W_3\right] + \mathrm{Cov}\left[W_1 W_2, W_3 W_4\right]\\ = 2\,\mathrm{Var}\left[\tau_1^{(n)}\right] - 4\,\mathrm{Cov}\left[\tau_1^{(n)}, \tau_2^{(n)}\right] + 3\,\mathrm{Cov}\left[\tau_1^{(n)}, \tau_3^{(n)}\right].\end{aligned}$$

In view of Lemmata 3.1 and 3.2 which are proven in Section 6, to verify (5) it suffices to show the following asymptotic result.

Proposition 3.1 *Consider the coalescent times* (7)*. As* $n \to \infty$*,*

$$\begin{aligned}\mathrm{Var}\left[\tau_1^{(n)}\right] &= \frac{\pi^2}{6} + O(n^{-1}\log^2 n),\\ \mathrm{Cov}\left[\tau_1^{(n)}, \tau_2^{(n)}\right] &= 2 - \frac{\pi^2}{6} + O(n^{-1}\log^2 n),\\ \mathrm{Cov}\left[\tau_1^{(n)}, \tau_3^{(n)}\right] &= 3 - \frac{5\pi^2}{18} + O(n^{-1}\log^2 n).\end{aligned}$$

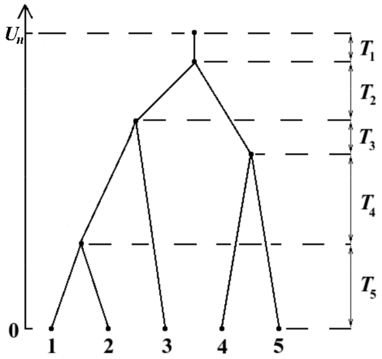

Figure 3: An example of a standard Yule n-tree with $n = 5$. The tree height is $U_n = T_1 + \ldots + T_n$, where T_i are the times between the consecutive speciation events. The 10 pairwise coalescent times $\tau_{ij}^{(n)}$ for the tips of the tree are $\tau_{12}^{(5)} = T_5$, $\tau_{45}^{(5)} = T_5 + T_4$, $\tau_{13}^{(5)} = \tau_{23}^{(5)} = T_5 + T_4 + T_3$, and $\tau_{14}^{(5)} = \tau_{24}^{(5)} = \tau_{34}^{(5)} = \tau_{15}^{(5)} = \tau_{25}^{(5)} = \tau_{35}^{(5)} = T_5 + T_4 + T_3 + T_2$.

Notice that the key Proposition 3.1 concerns only the first component of the evolutionary model we study - the standard Yule n-tree. For the standard Yule n-tree it is well known that the times between the consecutive speciation events $(T_1, \ldots, T_n)$ are independent exponentials with parameters $(1, \ldots, n)$ respectively, see Fig. 3. As shown in Gernhard [2008], this property corresponds to the unit rate Yule process conditioned on having n tips at the moment of observation, assuming that the time to the origin has the improper uniform prior [see also Feller, 1971].

4 Coalescent indices of the standard Yule n-tree

Following the standard Yule n-tree from its root toward the tips we label the consecutive splittings by indices $1, \ldots, n-1$: splitting k is the vertex when $k-1$ branches turn into k branches. We define three random splitting indices (as we interested in four randomly chosen tips out of n available):

- K_n is the index of the splitting where two randomly chosen tips coalesce,
- L_n be the index of the splitting where the first coalescent among three randomly chosen tips takes place,
- M_n be the index of the splitting where the first coalescent among four randomly chosen tips takes place.

To avoid multilevel indices in the forthcoming formulae, we will often use the following notational convention

$$KL_n := K_{L_n}, \quad LM_n := L_{M_n}, \quad KLM_n := K_{LM_n}.$$

To illustrate these indices, turn to the Fig. 3. If the two randomly chosen tips are $(1, 2)$, then $K_n = 4$. If the three randomly chosen tips are $(2, 3, 4)$, then $L_n = 4$, $K_{L_n} = 2$. If the four randomly chosen tips are $(2, 3, 4, 5)$, then $M_n = 3$, $L_{M_n} = 2$, $K_{LM_n} = 1$.

The importance of these random indices comes from the following representations. Denote $U_k^{(n)} := T_{k+1} + \ldots + T_n$ the sum of adjacent times between splittings in the Yule tree. Clearly,

$$\tau_1^{(n)} \overset{d}{=} U_{K_n}^{(n)}, \quad \tau_1^{(n)} \wedge \tau_2^{(n)} \overset{d}{=} U_{L_n}^{(n)}, \quad \tau_1^{(n)} \vee \tau_2^{(n)} \overset{d}{=} U_{KL_n}^{(n)}, \quad \tau_1^{(n)} \vee \tau_3^{(n)} \overset{d}{=} U_{KLM_n}^{(n)}. \tag{8}$$

To prove Proposition 3.1 we need to know the distributions of these random splitting indices. The next two lemmata giving these distributions are proved in Section 7.

Lemma 4.1 *Then*

$$\begin{aligned}
P(K_n = k) &= \frac{n+1}{n-1} \cdot \frac{2}{(k+1)(k+2)}, \quad k = 1, \ldots, n-1, \\
P(L_n = k) &= \frac{(n+1)(n+2)}{(n-1)(n-2)} \cdot \frac{6(k-1)}{(k+1)(k+2)(k+3)}, \quad k = 2, \ldots, n-1, \\
P(M_n = k) &= \frac{(n+1)(n+2)(n+3)}{(n-1)(n-2)(n-3)} \cdot \frac{12(k-1)(k-2)}{(k+1)(k+2)(k+3)(k+4)}, \quad k = 3, \ldots, n-1.
\end{aligned}$$

Lemma 4.2 *The random numbers $K_{L_n}, L_{M_n}, K_{LM_n}$ have the following distributions*

$$\begin{aligned}
P(K_{L_n} = k) &= \frac{n+1}{(n-1)(n-2)} \cdot \frac{12}{(k+1)(k+2)} \Big(\frac{n+2}{k+3} - 1\Big), \quad k = 1, \ldots, n-2, \\
P(L_{M_n} = k) &= \frac{(n+1)(n+2)}{(n-1)(n-2)(n-3)} \cdot \frac{72(k-1)}{(k+1)(k+2)(k+3)} \Big(\frac{n+3}{k+4} - 1\Big), \quad k = 2, \ldots, n-2, \\
P(K_{LM_n} = k) &= \frac{n+3}{(n-1)(n-2)(n-3)} \cdot \frac{72}{(k+1)(k+2)} \Big(\frac{(n+1)(n+2)}{(k+3)(k+4)} - 1\Big) \\
&\quad - \frac{n+2}{(n-1)(n-2)(n-3)} \cdot \frac{144}{(k+1)(k+2)} \Big(\frac{n+1}{k+3} - 1\Big), \quad k = 1, \ldots, n-3.
\end{aligned}$$

5 Proof of Proposition 3.1

In view of (8), the harmonic numbers

$$H_n = \sum_{k=1}^{n} k^{-1}, \qquad \bar{H}_n = \sum_{k=1}^{n} k^{-2}, \tag{9}$$

play an important role in our calculations as

$$\begin{aligned}
\mathrm{E}\Big[U_k^{(n)}\Big] &= H_n - H_k, \\
\mathrm{E}\Big[(U_k^{(n)})^2\Big] &= \bar{H}_n - \bar{H}_k + (H_n - H_k)^2 = H_n^2 - 2H_n H_k + \bar{H}_n + H_k^2 - \bar{H}_k.
\end{aligned}$$

Lemma 5.1 *We have*

$$
\begin{aligned}
\mathrm{E}\left[\tau_1^{(n)}\right] &= H_n - \mathrm{E}\left[H_{K_n}\right],\\
\mathrm{E}\left[(\tau_1^{(n)})^2\right] &= H_n^2 - 2H_n\,\mathrm{E}\left[H_{K_n}\right] + \bar{H}_n + \mathrm{E}\left[H_{K_n}^2 - \bar{H}_{K_n}\right],\\
\mathrm{E}\left[\tau_1^{(n)}\tau_2^{(n)}\right] &= H_n^2 - H_n\,\mathrm{E}\left[\frac{2H_{L_n}+4H_{KL_n}}{3}\right] + \bar{H}_n + \mathrm{E}\left[\frac{2H_{L_n}H_{KL_n}+H_{KL_n}^2-2\bar{H}_{L_n}-\bar{H}_{KL_n}}{3}\right],\\
\mathrm{E}\left[\tau_1^{(n)}\tau_3^{(n)}\right] &= H_n^2 - H_n\,\mathrm{E}\left[\frac{3H_{M_n}+5H_{LM_n}+10H_{KLM_n}}{9}\right] + \bar{H}_n - \frac{1}{3}\mathrm{E}\left[\bar{H}_{M_n}\right]\\
&\quad + \frac{1}{9}\mathrm{E}\left[H_{M_n}H_{LM_n} + 2H_{M_n}H_{KLM_n} + 4H_{LM_n}H_{KLM_n} + 2H_{KLM_n}^2 - 4\bar{H}_{LM_n} - 2\bar{H}_{KLM_n}\right].
\end{aligned}
$$

In view of Lemma 5.1, proven in Section 7, the asymptotic results stated in Proposition 3.1 are computed using the following relations involving the harmonic numbers (9).

Lemma 5.2 *We have as $n \to \infty$*

$$
\mathrm{E}\left[H_{K_n}\right] = 2 + O(n^{-1}\log n),\quad \mathrm{E}\left[H_{L_n}\right] = 3 + O(n^{-1}\log n),\quad \mathrm{E}\left[H_{M_n}\right] = \frac{11}{3} + O(n^{-1}\log n),
$$
$$
\mathrm{E}\left[H_{KL_n}\right] = \frac{3}{2} + O(n^{-1}\log n),\quad \mathrm{E}\left[H_{LM_n}\right] = \frac{7}{3} + O(n^{-1}\log n),\quad \mathrm{E}\left[H_{KLM_n}\right] = \frac{4}{3} + O(n^{-1}\log n).
$$

Lemma 5.3 *Let $a_n \rightrightarrows a$ stand for $a_n = a + O(n^{-1}\log^2 n)$ as $n \to \infty$. Then*

$$
\begin{array}{lll}
\mathrm{E}\left[H_{K_n}^2\right] \rightrightarrows \frac{\pi^2}{3}+2, & \mathrm{E}\left[H_{L_n}^2\right] \rightrightarrows \frac{21}{2}, & \mathrm{E}\left[H_{M_n}^2\right] \rightrightarrows \frac{\pi^2}{3}+\frac{211}{18},\\
\mathrm{E}\left[\bar{H}_{K_n}\right] \rightrightarrows \frac{\pi^2}{3}-2, & \mathrm{E}\left[\bar{H}_{L_n}\right] \rightrightarrows \frac{3}{2}, & \mathrm{E}\left[\bar{H}_{M_n}\right] \rightrightarrows \frac{\pi^2}{3}-\frac{31}{18},
\end{array}
$$

and

$$
\begin{array}{lll}
\mathrm{E}\left[H_{KL_n}^2\right] \rightrightarrows \frac{\pi^2}{2}-\frac{9}{4}, & \mathrm{E}\left[H_{LM_n}^2\right] \rightrightarrows \frac{167}{18}-\frac{\pi^2}{3}, & \mathrm{E}\left[H_{KLM_n}^2\right] \rightrightarrows \frac{2\pi^2}{3}-\frac{41}{9},\\
\mathrm{E}\left[\bar{H}_{KL_n}\right] \rightrightarrows \frac{\pi^2}{2}-\frac{15}{4}, & \mathrm{E}\left[\bar{H}_{LM_n}\right] \rightrightarrows \frac{85}{18}-\frac{\pi^2}{3}, & \mathrm{E}\left[\bar{H}_{KLM_n}\right] \rightrightarrows \frac{2\pi^2}{3}-\frac{49}{9},
\end{array}
$$

and

$$
\begin{array}{ll}
\mathrm{E}\left[H_{L_n}H_{KL_n}\right] \rightrightarrows \frac{39}{4}-\frac{\pi^2}{2}, & \mathrm{E}\left[H_{M_n}H_{LM_n}\right] \rightrightarrows \frac{221}{18}-\frac{\pi^2}{3},\\
\mathrm{E}\left[H_{M_n}H_{KLM_n}\right] \rightrightarrows \frac{2\pi^2}{3}-\frac{14}{9}, & \mathrm{E}\left[H_{LM_n}H_{KLM_n}\right] \rightrightarrows \frac{148}{9}-\frac{4\pi^2}{3}.
\end{array}
$$

The proofs of the last two lemmata are given in Section 8 using the auxiliary results from Appendix A.

With Lemmata 5.1 - 5.3 at hand, the remaining proof of Proposition 3.1 is straightforward. The first statement

$$
\mathrm{Var}\left[\tau_1^{(n)}\right] = \mathrm{E}\left[(\tau_1^{(n)})^2\right] - \left(\mathrm{E}\left[\tau_1^{(n)}\right]\right)^2 = \bar{H}_n + \mathrm{E}\left[H_{K_n}^2 - \bar{H}_{K_n}\right] - (\mathrm{E}\left[H_{K_n}\right])^2 \rightrightarrows \frac{\pi^2}{6},
$$

is obtained applying the classical relation $\bar{H}_n = \frac{\pi^2}{6} + O(n^{-1})$. Further, Lemma 5.1 yields

$$\operatorname{Cov}\left[\tau_1^{(n)}, \tau_2^{(n)}\right] = \mathrm{E}\left[\tau_1^{(n)}\tau_2^{(n)}\right] - \left(\mathrm{E}\left[\tau_1^{(n)}\right]\right)^2 = H_n\,\mathrm{E}\left[2H_{K_n} - \frac{2H_{L_n} + 4H_{KL_n}}{3}\right] \\ + \bar{H}_n + \frac{1}{3}\mathrm{E}\left[2H_{L_n}H_{KL_n} + H_{KL_n}^2 - 2\bar{H}_{L_n} - \bar{H}_{KL_n}\right] - (\mathrm{E}\left[H_{K_n}\right])^2,$$

where according to Lemma 5.2

$$\mathrm{E}\left[2H_{K_n} - \frac{2H_{L_n} + 4H_{KL_n}}{3}\right] = O(n^{-1}\log n).$$

Thus, applying Lemma 5.3 we obtain the second statement

$$\operatorname{Cov}\left[\tau_1^{(n)}, \tau_2^{(n)}\right] \rightrightarrows \frac{\pi^2}{6} + \frac{1}{3}\left(+2\left(\frac{39}{4} - \frac{\pi^2}{2}\right) + \frac{\pi^2}{2} - \frac{9}{4} - 2\cdot\frac{3}{2} - \frac{\pi^2}{2} + \frac{15}{4}\right) - 4 = 2 - \frac{\pi^2}{6}.$$

Finally, the third statement follows from

$$\operatorname{Cov}\left[\tau_1^{(n)}, \tau_3^{(n)}\right] = H_n\,\mathrm{E}\left[2H_{K_n} - \frac{3H_{M_n} + 5H_{LM_n} + 10H_{KLM_n}}{9}\right] + \bar{H}_n - (\mathrm{E}\left[H_{K_n}\right])^2 - \frac{1}{3}\mathrm{E}\left[\bar{H}_{M_n}\right] \\ + \frac{1}{9}\mathrm{E}\left[H_{M_n}H_{LM_n} + 2H_{M_n}H_{KLM_n} + 4H_{LM_n}H_{KLM_n} + 2H_{KLM_n}^2 - 4\bar{H}_{LM_n} - 2\bar{H}_{KLM_n}\right].$$

Indeed, according to Lemma 5.2

$$\mathrm{E}\left[2H_{K_n} - \frac{3H_{M_n} + 5H_{LM_n} + 10H_{KLM_n}}{9}\right] = O(n^{-1}\log n).$$

Moreover, from the following three limits

$$\bar{H}_n - (\mathrm{E}\left[H_{K_n}\right])^2 - \frac{1}{3}\mathrm{E}\left[\bar{H}_{M_n}\right] \rightrightarrows \frac{\pi^2}{18} - \frac{185}{54},$$
$$\mathrm{E}\left[H_{M_n}H_{LM_n} + 2H_{M_n}H_{KLM_n} + 4H_{LM_n}H_{KLM_n}\right] \rightrightarrows \frac{1349}{18} - \frac{13\pi^2}{3},$$
$$\mathrm{E}\left[H_{KLM_n}^2 - 2\bar{H}_{LM_n} - \bar{H}_{KLM_n}\right] \rightrightarrows \frac{2\pi^2}{3} - \frac{77}{9},$$

we get the stated overall limit

$$\frac{\pi^2}{18} - \frac{185}{54} + \frac{1}{9}\left(\frac{1349}{18} - \frac{13\pi^2}{3}\right) + \frac{2}{9}\left(\frac{2\pi^2}{3} - \frac{77}{9}\right) = 3 - \frac{5\pi^2}{18}.$$

6 Proofs of Lemmata 3.1 - 3.2

PROOF OF LEMMA 3.1. Using the representation

$$D_n^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \overline{Y}_n^2\right) = \frac{1}{n}\sum_i Y_i^2 - \frac{1}{n(n-1)}\sum_i\sum_{j\neq i} Y_iY_j$$

we find that

$$\begin{aligned}
\mathrm{E}\left[D_n^4\right] = {} & \frac{1}{n^2}\Big(\sum_i \mathrm{E}\left[Y_i^4\right] + \sum_i\sum_{j\neq i}\mathrm{E}\left[Y_i^2Y_j^2\right]\Big) \\
& - \frac{2}{n^2(n-1)}\Big(\sum_i\sum_{j\neq i}\mathrm{E}\left[Y_i^3Y_j\right] + \sum_i\sum_{j\neq i}\mathrm{E}\left[Y_iY_j^3\right] + \sum_i\sum_{j\neq i}\sum_{k\neq i,j}\mathrm{E}\left[Y_i^2Y_jY_k\right]\Big) \\
& + \frac{1}{n^2(n-1)^2}\Big(\sum_i\sum_{j\neq i}\mathrm{E}\left[Y_i^2Y_j^2\right] + \sum_i\sum_{j\neq i}\sum_{k\neq i,j}\mathrm{E}\left[Y_i^2Y_jY_k\right] + \sum_i\sum_{j\neq i}\sum_{k\neq i,j}\mathrm{E}\left[Y_iY_j^2Y_k\right] \\
& \qquad + \sum_i\sum_{j\neq i}\sum_{k\neq i,j}\sum_{l\neq i,j,k}\mathrm{E}\left[Y_iY_jY_kY_l\right]\Big).
\end{aligned}$$

If (W_1,W_2,W_3,W_4) is a random sample without replacement of four out of n trait values, then

$$\begin{aligned}
\mathrm{E}\left[W_1^4\right] &= n^{-1}\sum_i \mathrm{E}\left[Y_i^4\right], \\
\mathrm{E}\left[W_1^3W_2\right] &= \frac{1}{n(n-1)}\sum_i\sum_{j\neq i}\mathrm{E}\left[Y_i^3Y_j\right], \\
\mathrm{E}\left[W_1^2W_2^2\right] &= \frac{1}{n(n-1)}\sum_i\sum_{j\neq i}\mathrm{E}\left[Y_i^2Y_j^2\right],
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}\left[W_1^2W_2W_3\right] &= \frac{1}{n(n-1)(n-2)}\sum_i\sum_{j\neq i}\sum_{k\neq i,j}\mathrm{E}\left[Y_i^2Y_jY_k\right], \\
\mathrm{E}\left[W_1W_2W_3W_4\right] &= \frac{1}{n(n-1)(n-2)(n-3)}\sum_i\sum_{j\neq i}\sum_{k\neq i,j}\sum_{l\neq i,j,k}\mathrm{E}\left[Y_iY_jY_kY_l\right].
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\mathrm{E}\left[D_n^4\right] = {} & n^{-1}\,\mathrm{E}\left[W_1^4\right] - 4n^{-1}\,\mathrm{E}\left[W_1^3W_2\right] + \frac{n^2-2n+2}{n(n-1)}\mathrm{E}\left[W_1^2W_2^2\right] \\
& - \frac{2(n-2)^2}{n(n-1)}\mathrm{E}\left[W_1^2W_2W_3\right] + \frac{(n-2)(n-3)}{n(n-1)}\mathrm{E}\left[W_1W_2W_3W_4\right].
\end{aligned} \tag{10}$$

Since

$$\mathrm{E}\left[D_n^2\right] = \mathrm{E}\left[W_1^2\right] - \mathrm{E}\left[W_1 W_2\right],$$

we conclude

$$\begin{aligned}\mathrm{Var}\left[D_n^2\right] = n^{-1}\,\mathrm{E}\left[W_1^4\right] - 4n^{-1}\,\mathrm{E}\left[W_1^3 W_2\right] + \mathrm{Cov}\left[W_1^2, W_2^2\right] - \frac{n-2}{n(n-1)}\,\mathrm{E}\left[W_1^2 W_2^2\right] \\ - 2\,\mathrm{Cov}\left[W_1^2, W_2 W_3\right] + \frac{2(3n-4)}{n(n-1)}\,\mathrm{E}\left[W_1^2 W_2 W_3\right] \\ + \mathrm{Cov}\left[W_1 W_2, W_3 W_4\right] - \frac{2(2n-3)}{n(n-1)}\,\mathrm{E}\left[W_1 W_2 W_3 W_4\right].\end{aligned}$$

The stated relations follow with

$$\begin{aligned}B_n = \mathrm{E}\left[W_1^4\right] - 4\,\mathrm{E}\left[W_1^3 W_2\right] \\ - \frac{n-2}{n-1}\,\mathrm{E}\left[W_1^2 W_2^2\right] + \frac{2(3n-4)}{n-1}\,\mathrm{E}\left[W_1^2 W_2 W_3\right] - \frac{2(2n-3)}{n-1}\,\mathrm{E}\left[W_1 W_2 W_3 W_4\right].\end{aligned}$$

□

PROOF OF LEMMA 3.2. Denote by $Y_{ij}^{(n)}$ the normalized trait value of the most recent common ancestor of the tips (i,j). Let $\mathscr{Y}_{ij}^{(n)}$ stand for the σ–algebra generated by the pair $(\mathscr{Y}_n, Y_{ij}^{(n)})$, then

$$\begin{aligned}\mathrm{E}\left[Y_i|\mathscr{Y}_{ij}^{(n)}\right] = \mathrm{E}\left[Y_j|\mathscr{Y}_{ij}^{(n)}\right] = Y_{ij}^{(n)}, \\ \mathrm{Var}\left[Y_i|\mathscr{Y}_{ij}^{(n)}\right] = \mathrm{Var}\left[Y_j|\mathscr{Y}_{ij}^{(n)}\right] = \tau_{ij}^{(n)}, \\ \mathrm{Cov}\left[Y_i, Y_j|\mathscr{Y}_{ij}^{(n)}\right] = 0,\end{aligned}$$

implying (6)

$$\mathrm{Cov}\left[Y_i, Y_j|\mathscr{Y}_n\right] = \mathrm{Var}\left[Y_{ij}^{(n)}|\mathscr{Y}_n\right] = U_n - \tau_{ij}^{(n)}.$$

By Eq. (13) of Bohrnstedt and Goldberger [1969], we have

$$\mathrm{Cov}\left[Z_i Z_j, Z_k Z_l\right] = m_i m_k c_{jl} + m_i m_l c_{jk} + m_j m_k c_{il} + m_j m_l c_{ik} + c_{ik} c_{jl} + c_{il} c_{jk}$$

for any sequence of normally distributed random values $Z_1, Z_2, \ldots$ with means $\mathrm{E}[Z_i] = m_i$ and covariances $\mathrm{Cov}\left[Z_i, Z_j\right] = c_{ij}$. In the special case with $m_i = 0$ it follows

$$\begin{aligned}\mathrm{Cov}\left[Z_i Z_j, Z_k Z_l\right] = c_{ik} c_{jl} + c_{il} c_{jk}, \\ \mathrm{Cov}\left[Z_i^2, Z_j Z_k\right] = 2 c_{ij} c_{ik}, \\ \mathrm{Cov}\left[Z_i^2, Z_j^2\right] = 2 c_{ij}^2.\end{aligned}$$

Using conditional normality of Y_i and putting $c_{ij} = U_n - \tau_{ij}^{(n)}$, we derive from these relations that

$$\begin{aligned}
\operatorname{Cov}\left[Y_i^2, Y_j^2 | \mathscr{Y}_n\right] &= 2(U_n - \tau_{ij}^{(n)})^2, \\
\operatorname{Cov}\left[Y_i^2, Y_jY_k | \mathscr{Y}_n\right] &= 2(U_n - \tau_{ij}^{(n)})(U_n - \tau_{ik}^{(n)}), \\
\operatorname{Cov}\left[Y_iY_j, Y_kY_l | \mathscr{Y}_n\right] &= (U_n - \tau_{ik}^{(n)})(U_n - \tau_{jl}^{(n)}) + (U_n - \tau_{il}^{(n)})(U_n - \tau_{jk}^{(n)}).
\end{aligned}$$

yielding in terms of (7),

$$\begin{aligned}
\operatorname{Cov}\left[W_1^2, W_2^2 | \mathscr{Y}_n\right] &= 2(U_n - \tau_1^{(n)})^2, \\
\operatorname{Cov}\left[W_1^2, W_2W_3 | \mathscr{Y}_n\right] &= 2(U_n - \tau_1^{(n)})(U_n - \tau_2^{(n)}), \\
\operatorname{Cov}\left[W_1W_2, W_3W_4 | \mathscr{Y}_n\right] &= (U_n - \tau_2^{(n)})(U_n - \tau_5^{(n)}) + (U_n - \tau_3^{(n)})(U_n - \tau_4^{(n)}).
\end{aligned}$$

By the total covariance formula, we derive

$$\begin{aligned}
\operatorname{Cov}\left[W_1^2, W_2^2\right] &= 2\operatorname{E}\left[(U_n - \tau_1^{(n)})^2\right] + \operatorname{Var}\left[U_n\right], \\
\operatorname{Cov}\left[W_1^2, W_2W_3\right] &= 2\operatorname{E}\left[(U_n - \tau_1^{(n)})(U_n - \tau_2^{(n)})\right] + \operatorname{Cov}\left[U_n, U_n - \tau_1^{(n)}\right], \\
\operatorname{Cov}\left[W_1W_2, W_3W_4\right] &= 2\operatorname{E}\left[(U_n - \tau_1^{(n)})(U_n - \tau_3^{(n)})\right] + \operatorname{Cov}\left[U_n - \tau_1^{(n)}, U_n - \tau_3^{(n)}\right] \\
&= 3\operatorname{E}\left[(U_n - \tau_1^{(n)})(U_n - \tau_3^{(n)})\right] - \left(\operatorname{E}\left[U_n - \tau_1^{(n)}\right]\right)^2.
\end{aligned}$$

Combining these relations we get

$$\begin{aligned}
&\operatorname{Cov}\left[W_1^2, W_2^2\right] - 2\operatorname{Cov}\left[W_1^2, W_2W_3\right] + \operatorname{Cov}\left[W_1W_2, W_3W_4\right] \\
&\quad = 2\operatorname{E}\left[(U_n - \tau_1^{(n)})^2\right] - 4\operatorname{E}\left[(U_n - \tau_1^{(n)})(U_n - \tau_2^{(n)})\right] \\
&\qquad + 3\operatorname{E}\left[(U_n - \tau_1^{(n)})(U_n - \tau_3^{(n)})\right] \\
&\qquad + \operatorname{Var}\left[U_n\right] - 2\operatorname{Cov}\left[U_n, U_n - \tau_1^{(n)}\right] - \left(\operatorname{E}\left[U_n - \tau_1^{(n)}\right]\right)^2.
\end{aligned}$$

This together with

$$\begin{aligned}
&\operatorname{Var}\left[U_n\right] - 2\operatorname{Cov}\left[U_n, U_n - \tau_1^{(n)}\right] - \left(\operatorname{E}\left[U_n - \tau_1^{(n)}\right]\right)^2 \\
&\quad = \operatorname{E}\left[U_n^2\right] - \operatorname{E}\left[U_n\right]^2 - 2\operatorname{E}\left[U_n(U_n - \tau_1^{(n)})\right] + 2\operatorname{E}\left[U_n\right]\operatorname{E}\left[U_n - \tau_1^{(n)}\right] - \left(\operatorname{E}\left[U_n - \tau_1^{(n)}\right]\right)^2 \\
&\quad = \operatorname{E}\left[U_n^2\right] - 2\operatorname{E}\left[U_n(U_n - \tau_1^{(n)})\right] - \operatorname{E}\left[\tau_1^{(n)}\right]^2 \\
&\quad = \operatorname{E}\left[(\tau_1^{(n)})^2\right] - \operatorname{E}\left[\tau_1^{(n)}\right]^2 - \operatorname{E}\left[(U_n - \tau_1^{(n)})^2\right]
\end{aligned}$$

implies the assertion of the Lemma 3.2

$$\begin{aligned}
\mathrm{Cov}\left[W_1^2, W_2^2\right] - 2\,\mathrm{Cov}\left[W_1^2, W_2W_3\right] + \mathrm{Cov}\left[W_1W_2, W_3W_4\right]& \\
= \mathrm{E}\left[(U_n - \tau_1^{(n)})^2\right] + \mathrm{E}\left[(\tau_1^{(n)})^2\right] - \mathrm{E}\left[\tau_1^{(n)}\right]^2& \\
- 4\,\mathrm{E}\left[(U_n - \tau_1^{(n)})(U_n - \tau_2^{(n)})\right] + 3\,\mathrm{E}\left[(U_n - \tau_1^{(n)})(U_n - \tau_3^{(n)})\right]& \\
= \mathrm{E}\left[(U_n - \tau_1^{(n)})^2\right] + \mathrm{Var}\left[\tau_1^{(n)}\right] - \mathrm{E}\left[(U_n - \tau_1^{(n)})U_n\right] + \mathrm{E}\left[U_n\tau_1^{(n)}\right]& \\
- 4\,\mathrm{E}\left[\tau_1^{(n)}\tau_2^{(n)}\right] + 3\,\mathrm{E}\left[\tau_1^{(n)}\tau_3^{(n)}\right]& \\
= 2\,\mathrm{Var}\left[\tau_1^{(n)}\right] - 4\,\mathrm{Cov}\left[\tau_1^{(n)}, \tau_2^{(n)}\right] + 3\,\mathrm{Cov}\left[\tau_1^{(n)}, \tau_3^{(n)}\right].&
\end{aligned}$$

□

7 Proofs of Lemmata 4.1, 4.2, and 5.1

PROOF of Lemma 4.1 From the definition of K_n it is easy to see that, for $k = 2, \ldots, n$,

$$P(K_n < k-1 | K_n < k) = 1 - \frac{1}{\binom{k}{2}} = \frac{(k+1)(k-2)}{k(k-1)}.$$

Therefore,

$$\begin{aligned}
P(K_n < k-1) &= \frac{(n+1)(n-2)}{n(n-1)} \frac{n(n-3)}{(n-1)(n-2)} \frac{(n-1)(n-4)}{(n-2)(n-3)} \cdots \frac{(k+1)(k-2)}{k(k-1)} \\
&= \frac{(n+1)(k-2)}{(n-1)k}, \\
P(K_n = k-1) &= \frac{(n+1)(k-1)}{(n-1)(k+1)} - \frac{(n+1)(k-2)}{(n-1)k} = \frac{n+1}{n-1} \frac{2}{(k+1)k}.
\end{aligned}$$

Similarly, for $k = 3, \ldots, n$,

$$\begin{aligned}
P(L_n < k-1 | L_n < k) &= 1 - \frac{3}{\binom{k}{2}} = \frac{(k+2)(k-3)}{k(k-1)}, \\
P(L_n < k-1) &= \frac{(n+2)(n-3)}{n(n-1)} \frac{(n+1)(n-4)}{(n-1)(n-2)} \frac{n(n-5)}{(n-2)(n-3)} \cdots \frac{(k+2)(k-3)}{k(k-1)} \\
&= \frac{(n+2)(n+1)(k-2)(k-3)}{(n-1)(n-2)(k+1)k}, \\
P(L_n = k-1) &= \frac{(n+2)(n+1)(k-1)(k-2)}{(n-1)(n-2)(k+2)(k+1)} - \frac{(n+2)(n+1)(k-2)(k-3)}{(n-1)(n-2)(k+1)k} \\
&= \frac{6(n+2)(n+1)(k-2)}{(n-1)(n-2)(k+2)(k+1)k},
\end{aligned}$$

and, for $k = 4, \dots, n$,

$$P(M_n < k-1 | M_n < k) = 1 - \frac{6}{\binom{k}{2}} = \frac{(k+3)(k-4)}{k(k-1)},$$
$$P(M_n < k-1) = \frac{(n+3)(n+2)(n+1)(k-2)(k-3)(k-4)}{(n-1)(n-2)(n-3)(k+2)(k+1)k},$$
$$P(M_n = k-1) = \frac{12(n+3)(n+2)(n+1)(k-2)(k-3)}{(n-1)(n-2)(n-3)(k+3)(k+2)(k+1)k}.$$

□

PROOF of Lemma 4.2 Clearly,

$$\begin{aligned} P(K_{L_n} = k) &= \sum_{l=k+1}^{n-1} P(K_l = k) P(L_n = l) \\ &= \frac{(n+1)(n+2)}{(n-1)(n-2)} \cdot \frac{12}{(k+1)(k+2)} \sum_{l=k+1}^{n-1} \frac{1}{(l+2)(l+3)} \end{aligned}$$

leads to the first assertion. Further,

$$\begin{aligned} P(L_{M_n} = k) = &\frac{(n+1)(n+2)(n+3)}{(n-1)(n-2)(n-3)} \\ &\times \sum_{m=k+1}^{n-1} \frac{12(m-1)(m-2)(m+1)(m+2)}{(m+1)(m+2)(m+3)(m+4)(m-1)(m-2)} \cdot \frac{6(k-1)}{(k+1)(k+2)(k+3)}, \end{aligned}$$

and

$$\begin{aligned} P(L_{M_n} = k) &= \frac{(n+1)(n+2)(n+3)}{(n-1)(n-2)(n-3)} \cdot \frac{72(k-1)}{(k+1)(k+2)(k+3)} \sum_{m=k+1}^{n-1} \frac{1}{(m+3)(m+4)} \\ &= \frac{(n+1)(n+2)(n+3)}{(n-1)(n-2)(n-3)} \cdot \frac{72(k-1)}{(k+1)(k+2)(k+3)} (\frac{1}{k+4} - \frac{1}{n+3}). \end{aligned}$$

Finally,

$$\begin{aligned} P(K_{LM_n} = k) = &\sum_{m=k+1}^{n-2} \frac{(n+1)(n+2)(n+3)}{(n-1)(n-2)(n-3)} \cdot \frac{72(m-1)}{(m+1)(m+2)(m+3)(m+4)} \frac{m+1}{m-1} \cdot \frac{2}{(k+1)(k+2)} \\ &- \sum_{m=k+1}^{n-2} \frac{(n+1)(n+2)}{(n-1)(n-2)(n-3)} \cdot \frac{72(m-1)}{(m+1)(m+2)(m+3)} \frac{m+1}{m-1} \cdot \frac{2}{(k+1)(k+2)} \\ = &\frac{(n+1)(n+2)(n+3)}{(n-1)(n-2)(n-3)} \cdot \frac{144}{(k+1)(k+2)} \sum_{m=k+1}^{n-2} \frac{1}{(m+2)(m+3)(m+4)} \\ &- \frac{(n+1)(n+2)}{(n-1)(n-2)(n-3)} \cdot \frac{144}{(k+1)(k+2)} \sum_{m=k+1}^{n-2} \frac{1}{(m+2)(m+3)}, \end{aligned}$$

and therefore, it remains to use the equalities

$$\sum_{m=k+1}^{n-2} \frac{1}{(m+2)(m+3)} = \frac{1}{k+3} - \frac{1}{n+1},$$
$$\sum_{m=k+1}^{n-2} \frac{2}{(m+2)(m+3)(m+4)} = \frac{1}{(k+3)(k+4)} - \frac{1}{(n+1)(n+2)}.$$

□

PROOF OF LEMMA 5.1. We have

$$\mathrm{E}\left[\tau_1^{(n)}\right] = \mathrm{E}\left[U_{K_n}^{(n)}\right] = H_n - \mathrm{E}\left[H_{K_n}\right],$$

and

$$\mathrm{E}\left[(\tau_1^{(n)})^2\right] = \mathrm{E}\left[(U_{K_n}^{(n)})^2\right] = \bar{H}_n + H_n^2 - \mathrm{E}\left[\bar{H}_{K_n} + 2H_{K_n}H_n - H_{K_n}^2\right].$$

Further, using (8) and $U_{KL_n}^{(n)} = U_{L_n}^{(n)} + U_{KL_n}^{(L_n)}$, we get

$$\begin{aligned}
\mathrm{E}\left[\tau_1^{(n)}\tau_2^{(n)}\right] &= \frac{1}{3}\mathrm{E}\left[(U_{KL_n}^{(n)})^2\right] + \frac{2}{3}\mathrm{E}\left[U_{KL_n}^{(n)}U_{L_n}^{(n)}\right] \\
&= \mathrm{E}\left[(U_{L_n}^{(n)})^2\right] + \frac{4}{3}\mathrm{E}\left[U_{L_n}^{(n)}U_{KL_n}^{(L_n)}\right] + \frac{1}{3}\mathrm{E}\left[(U_{KL_n}^{(L_n)})^2\right].
\end{aligned}$$

Thus

$$\begin{aligned}
\mathrm{E}\left[\tau_1^{(n)}\tau_2^{(n)}\right] &= H_n^2 - 2H_n\mathrm{E}\left[H_{L_n}\right] + \bar{H}_n + \mathrm{E}\left[H_{L_n}^2 - \bar{H}_{L_n}\right] + \frac{4}{3}\mathrm{E}\left[(H_n - H_{L_n})(H_{L_n} - H_{KL_n})\right] \\
&\quad + \frac{1}{3}\mathrm{E}\left[H_{L_n}^2 - 2H_{L_n}H_{KL_n} + \bar{H}_{L_n} + H_{KL_n}^2 - \bar{H}_{KL_n}\right] \\
&= H_n^2 - H_n\mathrm{E}\left[\frac{2H_{L_n} + 4H_{KL_n}}{3}\right] + \bar{H}_n + \mathrm{E}\left[\frac{2H_{L_n}H_{KL_n} + H_{KL_n}^2 - 2\bar{H}_{L_n} - \bar{H}_{KL_n}}{3}\right].
\end{aligned}$$

Finally, for two pairs of sampled tips, we have three coalescent events to consider: going from four to three selected nodes, 4 → 3, going from three to two selected nodes, 3 → 2, and going from two to one selected nodes, 2 → 1. The coalescent 4 → 3 holds across the two pairs with probability $\frac{4}{\binom{4}{2}} = \frac{2}{3}$ and within a pair with probability $\frac{1}{3}$. Given the former outcome, the coalescent 3 → 2 holds again across the pairs with probability $\frac{1}{3}$ and within a pair with probability $\frac{2}{3}$. Otherwise, the coalescent 3 → 2 holds across the pairs with probability $\frac{2}{3}$ and within the second pair with probability $\frac{1}{3}$. The four possibilities ($\frac{2}{3} \times \frac{1}{3}$, $\frac{2}{3} \times \frac{2}{3}$, $\frac{1}{3} \times \frac{2}{3}$, $\frac{1}{3} \times \frac{1}{3}$) produce the following four terms in

$$\mathrm{E}\left[\tau_1^{(n)}\tau_3^{(n)}\right] = \frac{2}{9}\mathrm{E}\left[(U_{KLM_n}^{(n)})^2\right] + \frac{4}{9}\mathrm{E}\left[U_{LM_n}^{(n)}U_{KLM_n}^{(n)}\right] + \frac{2}{9}\mathrm{E}\left[U_{M_n}^{(n)}U_{KLM_n}^{(n)}\right] + \frac{1}{9}\mathrm{E}\left[U_{M_n}^{(n)}U_{LM_n}^{(n)}\right].$$

It follows,

$$\mathrm{E}\left[\tau_1^{(n)}\tau_3^{(n)}\right]=\mathrm{E}\left[(U_{M_n}^{(n)})^2+\frac{2}{9}(U_{KLM_n}^{(M_n)})^2+\frac{10}{9}U_{M_n}^{(n)}U_{KLM_n}^{(M_n)}+\frac{4}{9}U_{LM_n}^{(M_n)}U_{KLM_n}^{(M_n)}+\frac{5}{9}U_{M_n}^{(n)}U_{LM_n}^{(M_n)}\right].$$

Using the representation for $\mathrm{E}\left[\tau_1^{(M_n)}\tau_2^{(M_n)}\right]$,

$$\begin{aligned}&\frac{1}{3}\mathrm{E}\left[(U_{KLM_n}^{(M_n)})^2\right]+\frac{2}{3}\mathrm{E}\left[U_{KLM_n}^{(M_n)}U_{LM_n}^{(M_n)}\right]\\&\quad=\mathrm{E}\left[H_{M_n}^2-H_{M_n}\frac{2H_{LM_n}+4H_{KLM_n}}{3}+\bar{H}_{M_n}+\frac{2H_{LM_n}H_{KLM_n}+H_{KLM_n}^2-2\bar{H}_{LM_n}-\bar{H}_{KLM_n}}{3}\right],\end{aligned}$$

we can write

$$\begin{aligned}\mathrm{E}\left[\tau_1^{(n)}\tau_3^{(n)}\right]&=H_n^2-\mathrm{E}\left[2H_nH_{M_n}+\bar{H}_{M_n}-H_{M_n}^2\right]+\bar{H}_n+\frac{2}{3}\mathrm{E}\left[H_{M_n}^2-H_{M_n}\frac{2H_{LM_n}+4H_{KLM_n}}{3}+\bar{H}_{M_n}\right]\\&\quad+\frac{2}{3}\mathrm{E}\left[\frac{2H_{LM_n}H_{KLM_n}+H_{KLM_n}^2-2\bar{H}_{LM_n}-\bar{H}_{KLM_n}}{3}\right]\\&\quad+\frac{5}{9}\mathrm{E}\left[(H_n-H_{M_n})(H_{M_n}-H_{LM_n})\right]+\frac{10}{9}\mathrm{E}\left[(H_n-H_{M_n})(H_{M_n}-H_{KLM_n})\right],\end{aligned}$$

which after a rearrangement gives the last statement.

□

8 Proof of Lemmata 5.2 - 5.3

In this section we will often use the elementary relations of the following type

$$\frac{6}{(k+1)(k+2)(k+3)(k+4)}=\frac{1}{(k+1)(k+2)}-\frac{2}{(k+2)(k+3)}+\frac{1}{(k+3)(k+4)},\tag{11}$$

$$\frac{(k-1)(k-2)}{(k+1)(k+2)(k+3)(k+4)}=\frac{1}{(k+1)(k+2)}-\frac{5}{(k+2)(k+3)}+\frac{5}{(k+3)(k+4)},\tag{12}$$

$$\frac{6k}{(k+1)(k+2)(k+3)(k+4)}=-\frac{1}{(k+1)(k+2)}+\frac{5}{(k+2)(k+3)}-\frac{4}{(k+3)(k+4)},\tag{13}$$

$$\frac{k(k-5)}{(k+1)(k+2)(k+3)(k+4)}=\frac{1}{(k+1)(k+2)}+\frac{6}{(k+3)(k+4)}-\frac{6}{(k+2)(k+3)},\tag{14}$$

valid for all $k\geq 1$.

PROOF of Lemma 5.2. The first three stated relations are obtained using Lemmata 4.1 and A.1.

Equalities

$$\begin{aligned}
\mathrm{E}\left[H_{K_n}\right] &= \frac{n+1}{n-1}\sum_{k=1}^{n-1}\frac{2H_k}{(k+1)(k+2)} = \frac{2(n-H_n)}{n-1},\\
\mathrm{E}\left[H_{L_n}\right] &= \frac{(n+1)(n+2)}{(n-1)(n-2)}\sum_{k=2}^{n-1}\frac{6(k-1)H_k}{(k+1)(k+2)(k+3)}\\
&= \frac{6(n+1)(n+2)}{(n-1)(n-2)}\sum_{k=1}^{n-1}\Big(\frac{2H_k}{(k+2)(k+3)} - \frac{H_k}{(k+1)(k+2)}\Big) = \frac{3n(n+1)-6nH_n}{(n-1)(n-2)},
\end{aligned}$$

give the first and the second stated relations, and the third one follows from

$$\begin{aligned}
\mathrm{E}\left[H_{M_n}\right] &= \frac{(n+1)(n+2)(n+3)}{(n-1)(n-2)(n-3)}\sum_{k=3}^{n-1}\frac{12(k-1)(k-2)H_k}{(k+1)(k+2)(k+3)(k+4)}\\
&\overset{(12)}{=} \frac{12(n+1)(n+2)(n+3)}{(n-1)(n-2)(n-3)}\sum_{k=1}^{n-1}\Big(\frac{H_k}{(k+1)(k+2)} - \frac{5H_k}{(k+2)(k+3)} + \frac{5H_k}{(k+3)(k+4)}\Big)\\
&= \frac{\frac{11}{3}n^3 + 30n^2 + \frac{391}{3}n - 20 - 12(n^2+1)H_n}{(n-1)(n-2)(n-3)}.
\end{aligned}$$

The second three stated relations are obtained similarly using Lemmata 4.2 and A.1. Indeed,

$$\begin{aligned}
\mathrm{E}\left[H_{KL_n}\right] &= \frac{(n+1)(n+2)}{(n-1)(n-2)}\sum_{k=1}^{n-1}\frac{12H_k}{(k+1)(k+2)(k+3)} - \frac{n+1}{(n-1)(n-2)}\sum_{k=1}^{n-1}\frac{12H_k}{(k+1)(k+2)}\\
&= \frac{6(n+1)(n+2)}{(n-1)(n-2)}\Big(\sum_{k=1}^{n-1}\frac{H_k}{(k+1)(k+2)} - \sum_{k=1}^{n-1}\frac{H_k}{(k+2)(k+3)}\Big) - \frac{12(n-H_n)}{(n-1)(n-2)}\\
&= \frac{6n(n-H_n)}{(n-1)(n-2)} - \frac{6(3n^2+5n-4(n+1)H_n)}{4(n-1)(n-2)} = \frac{6H_n}{(n-1)(n-2)} + \frac{3n(n-5)}{2(n-1)(n-2)},
\end{aligned}$$

implying $\mathrm{E}\left[H_{KL_n}\right] = \frac{3}{2} + O(n^{-1}\log n)$. Furthermore,

$$\begin{aligned}
\mathrm{E}\left[H_{LM_n}\right] &= \sum_{k=2}^{\infty}\frac{72(k-1)H_k}{(k+1)(k+2)(k+3)(k+4)} + O(n^{-1}\log n) \overset{(11)(13)}{=} \frac{7}{3} + O(n^{-1}\log n),\\
\mathrm{E}\left[H_{KLM_n}\right] &= \sum_{k=1}^{\infty}\frac{72H_k}{(k+1)(k+2)(k+3)(k+4)} + O(n^{-1}\log n) \overset{(11)}{=} \frac{4}{3} + O(n^{-1}\log n).
\end{aligned}$$

□

PROOF of Lemma 5.3. The stated relations are obtained using Lemmas 4.1, 4.2, A.1, A.2. Firstly,

$$\begin{aligned}
\mathrm{E}\left[H_{K_n}^2\right] &= \frac{n+1}{n-1}\sum_{k=1}^{n-1}\frac{2H_k^2}{(k+1)(k+2)} = \frac{2(n+1)}{n-1}\Big(\bar{H}_n + \frac{n-H_n^2-2H_n}{n+1}\Big)\\
&= \frac{2(\bar{H}_n(n+1)+n-H_n^2-2H_n)}{n-1} \rightrightarrows \frac{\pi^2}{3}+2.
\end{aligned}$$

Similarly, we have

$$\mathrm{E}\left[\bar{H}_{K_n}\right] = \frac{n+1}{n-1}\sum_{k=1}^{n-1}\frac{2\bar{H}_k}{(k+1)(k+2)} = \frac{2(n+1)}{n-1}\cdot\frac{n\bar{H}_n - n}{n+1} = \frac{2n\bar{H}_n - 2n}{n-1} \rightrightarrows \frac{\pi^2}{3} - 2.$$

Observe that the limit is $\sum_{k=1}^{\infty}\frac{2\bar{H}_k}{(k+1)(k+2)}$. In the same manner we obtain

$$\mathrm{E}\left[H_{KL_n}^2\right] \rightrightarrows \sum_{k=1}^{\infty}\frac{12H_k^2}{(k+1)(k+2)(k+3)} = \sum_{k=1}^{\infty}\frac{6H_k^2}{(k+1)(k+2)} - \sum_{k=1}^{\infty}\frac{6H_k^2}{(k+2)(k+3)} = \frac{\pi^2}{2} - \frac{9}{4},$$
$$\mathrm{E}\left[\bar{H}_{KL_n}\right] \rightrightarrows \sum_{k=1}^{\infty}\frac{12\bar{H}_k}{(k+1)(k+2)(k+3)} = \sum_{k=1}^{\infty}\frac{6\bar{H}_k}{(k+1)(k+2)} - \sum_{k=1}^{\infty}\frac{6\bar{H}_k}{(k+2)(k+3)} = \frac{\pi^2}{2} - \frac{15}{4}.$$

Using the decomposition (12) we find

$$\mathrm{E}\left[H_{M_n}^2\right] \rightrightarrows \sum_{k=3}^{\infty}\frac{12(k-1)(k-2)H_k^2}{(k+1)(k+2)(k+3)(k+4)} = \frac{\pi^2}{3} + \frac{211}{18},$$
$$\mathrm{E}\left[\bar{H}_{M_n}\right] \rightrightarrows \sum_{k=3}^{\infty}\frac{12(k-1)(k-2)\bar{H}_k}{(k+1)(k+2)(k+3)(k+4)} = \frac{\pi^2}{3} - \frac{31}{18}.$$

Using the difference between (13) and (11) we find

$$\mathrm{E}\left[H_{LM_n}^2\right] \rightrightarrows \sum_{k=1}^{\infty}\frac{72(k-1)H_k^2}{(k+1)(k+2)(k+3)(k+4)} = \frac{167}{18} - \frac{\pi^2}{3},$$
$$\mathrm{E}\left[\bar{H}_{LM_n}\right] \rightrightarrows \sum_{k=1}^{\infty}\frac{72(k-1)\bar{H}_k}{(k+1)(k+2)(k+3)(k+4)} = \frac{85}{18} - \frac{\pi^2}{3}.$$

Using (11) we find

$$\mathrm{E}\left[H_{KLM_n}^2\right] \rightrightarrows \sum_{k=1}^{\infty}\frac{72H_k^2}{(k+1)(k+2)(k+3)(k+4)} = \frac{2\pi^2}{3} - \frac{41}{9},$$
$$\mathrm{E}\left[\bar{H}_{KLM_n}\right] \rightrightarrows \sum_{k=1}^{\infty}\frac{72\bar{H}_k}{(k+1)(k+2)(k+3)(k+4)} = \frac{2\pi^2}{3} - \frac{49}{9}.$$

Since $\mathrm{E}\left[H_{K_n}\right] = \frac{2(n-H_n)}{n-1}$, we have

$$\begin{aligned}\mathrm{E}\left[H_{L_n}H_{KL_n}\right] &\rightrightarrows \sum_{m=2}^{\infty}\frac{6(m-1)H_m}{(m+1)(m+2)(m+3)}\frac{2(m-H_m)}{m-1}\\ &= \sum_{m=1}^{\infty}\frac{12mH_m}{(m+1)(m+2)(m+3)} - \sum_{m=1}^{\infty}\frac{12H_m^2}{(m+1)(m+2)(m+3)}\\ &= \frac{15}{2} - (\frac{\pi^2}{2} - \frac{9}{4}) = \frac{39}{4} - \frac{\pi^2}{2},\end{aligned}$$

where we use the following corollary of Lemma A.1

$$\sum_{m=1}^{\infty}\frac{2mH_m}{(m+1)(m+2)(m+3)}=\sum_{m=1}^{\infty}\frac{3H_m}{(m+2)(m+3)}-\sum_{m=1}^{\infty}\frac{H_m}{(m+1)(m+2)}=\frac{5}{4}.$$

Similarly,

$$\begin{aligned}\mathrm{E}\left[H_{LM_n}H_{KLM_n}\right]&\rightrightarrows\sum_{m=2}^{\infty}\frac{72(m-1)H_m}{(m+1)(m+2)(m+3)(m+4)}\frac{2(m-H_m)}{m-1}\\&=\sum_{m=1}^{\infty}\frac{144mH_m}{(m+1)(m+2)(m+3)(m+4)}-\sum_{m=1}^{\infty}\frac{144H_m^2}{(m+1)(m+2)(m+3)(m+4)},\end{aligned}$$

where

$$\begin{aligned}\sum_{k=1}^{\infty}\frac{144kH_k}{(k+1)(k+2)(k+3)(k+4)}&\overset{(13)}{=}24(-1+15/4-22/9)=\frac{22}{3},\\\sum_{k=1}^{\infty}\frac{144H_k^2}{(k+1)(k+2)(k+3)(k+4)}&\overset{(11)}{=}\frac{4\pi^2}{3}-\frac{82}{9},\end{aligned}$$

so that $\mathrm{E}\left[H_{LM_n}H_{KLM_n}\right]\rightrightarrows\frac{148}{9}-\frac{4\pi^2}{3}$.

Further, in view of $\mathrm{E}\left[H_{L_n}\right]=\frac{3n(n+1)-6nH_n}{(n-1)(n-2)}$, the limit for $\mathrm{E}\left[H_{M_n}H_{LM_n}\right]$ can be computed as

$$\begin{aligned}\mathrm{E}\left[H_{M_n}H_{LM_n}\right]&\rightrightarrows\sum_{m=1}^{\infty}\frac{12(m-1)(m-2)H_m}{(m+1)(m+2)(m+3)(m+4)}\frac{3m(m+1)-6mH_m}{(m-1)(m-2)}\\&=\sum_{m=1}^{\infty}\frac{36mH_m}{(m+2)(m+3)(m+4)}-\sum_{m=1}^{\infty}\frac{72mH_m^2}{(m+1)(m+2)(m+3)(m+4)},\end{aligned}$$

where

$$\begin{aligned}\sum_{m=1}^{\infty}\frac{6mH_m^2}{(m+1)(m+2)(m+3)(m+4)}&\overset{(13)}{=}\frac{\pi^2}{36}+\frac{85}{216},\\\sum_{m=1}^{\infty}\frac{mH_m}{(m+2)(m+3)(m+4)}&=\sum_{m=1}^{\infty}\frac{2H_m}{(m+3)(m+4)}-\sum_{m=1}^{\infty}\frac{H_m}{(m+2)(m+3)}=\frac{17}{36},\end{aligned}$$

yielding $\mathrm{E}\left[H_{M_n}H_{LM_n}\right]\rightrightarrows\frac{221}{18}-\frac{\pi^2}{3}$. Finally, from

$$\mathrm{E}\left[H_{KL_m}\right]=\frac{6H_m}{(m-1)(m-2)}+\frac{3m(m-5)}{2(m-1)(m-2)}$$

we get

$$\begin{aligned}\mathrm{E}\left[H_{M_n}H_{KLM_n}\right]&\rightrightarrows\sum_{m=1}^{\infty}\frac{12(m-1)(m-2)H_m}{(m+1)(m+2)(m+3)(m+4)}\left(\frac{6H_m}{(m-1)(m-2)}+\frac{3m(m-5)}{2(m-1)(m-2)}\right)\\&=\sum_{m=1}^{\infty}\frac{72H_m^2}{(m+1)(m+2)(m+3)(m+4)}+\sum_{m=1}^{\infty}\frac{18m(m-5)H_m}{(m+1)(m+2)(m+3)(m+4)},\end{aligned}$$

where

$$\begin{aligned}\sum_{m=1}^{\infty}\frac{72H_m^2}{(m+1)(m+2)(m+3)(m+4)} &\overset{(11)}{=} \frac{2\pi^2}{3}-\frac{41}{9},\\ \sum_{m=1}^{\infty}\frac{m(m-5)H_m}{(m+1)(m+2)(m+3)(m+4)} &\overset{(14)}{=} \frac{1}{6},\end{aligned}$$

so that $\mathrm{E}\left[H_{M_n}H_{KLM_n}\right] \rightrightarrows \frac{2\pi^2}{3}-\frac{14}{9}$.

□

Acknowledgements

The research of Serik Sagitov was supported by the Swedish Research Council grant 621-2010-5623. Krzysztof Bartoszek was supported by the Centre for Theoretical Biology at the University of Gothenburg, Svenska Institutets Östersjösamarbete scholarship nr. 11142/2013, Stiftelsen för Vetenskaplig Forskning och Utbildning i Matematik (Foundation for Scientific Research and Education in Mathematics), Knut and Alice Wallenbergs travel fund, Paul and Marie Berghaus fund, the Royal Swedish Academy of Sciences, and Wilhelm and Martina Lundgrens research fund.

References

V. Adamchik. On Stirling numbers and Euler sums. *J. Comput. Appl. Math.*, 79(1):119–130, 1997.

D. Aldous and L. Popovic. A critical branching process model for biodiversity. *Adv. Appl. Probab.*, 37(4):1094–1115, 2005.

C. Ané. Analysis of comparative data with hierarchical autocorrelation. *Ann. Appl. Stat*, 2(3): 1078–1102, 2008.

C. Ané, L. S. T. Ho, and S. Roch. Phase transition on the convergence rate of parameter estimation under an Ornstein–Uhlenbeck diffusion on a tree. *ArXiv e-prints*, 2014.

K. Bartoszek. Quantifying the effects of anagenetic and cladogenetic evolution. *Mathematical Biosciences*, 254:42–57, 2014.

K. Bartoszek and S. Sagitov. Phylogenetic confidence intervals for the optimal trait value. *ArXiv e-prints*, July 2012.

K. Bartoszek, J. Pienaar, P. Mostad, S. Andersson, and T. F. Hansen. A phylogenetic comparative method for studying multivariate adaptation. *J. Theor. Biol.*, 314:204–215, 2012.

G. W. Bohrnstedt and A. S. Goldberger. On the exact covariance of products of random variables. *J. Am. Stat. Assoc.*, 64:1439–1442, 1969.

M. A. Butler and A. A. King. Phylogenetic comparative analysis: a modelling approach for adaptive evolution. *Am. Nat.*, 164(6):683–695, 2004.

F. W. Crawford and M. A. Suchard. Diversity, disparity, and evolutionary rate estimation for unresolved Yule trees. *Syst. Biol.*, 62(3):439–455, 2013.

A. W. F. Edwards. Estimation of the branch points of a branching diffusion process. *J. Roy. Stat. Soc. B*, 32(2):155–174, 1970.

W. Feller. *An Introduction to Probability Theory and Its Applications Vol. II.* John Wiley & Sons, New York, 1971.

J. Felsenstein. Phylogenies and the comparative method. *Am. Nat.*, 125(1):1–15, 1985.

J. Felsenstein. Phylogenies and quantitative characters. *Annu. Rev. Ecol. Syst.*, 19:445–471, 1988.

T. Gernhard. The conditioned reconstructed process. *J. Theor. Biol.*, 253:769–778, 2008.

T. F. Hansen. Stabilizing selection and the comparative analysis of adaptation. *Evolution*, 51(5): 1341–1351, 1997.

T. F. Hansen and K. Bartoszek. Interpreting the evolutionary regression: the interplay between observational and biological errors in phylogenetic comparative studies. *Syst. Biol.*, 61(3):413–425, 2012.

T. F. Hansen, J. Pienaar, and S. H. Orzack. A comparative method for studying adaptation to a randomly evolving environment. *Evolution*, 62:1965–1977, 2008.

K. M. Helgen, C. Miguel Pinto, R. Keys, L. E. Helgren, M. T. N. Tsuchiya, A. Quinn, D. E. Wilson, and J. E. Maldonado. Taxonomic revision of the olingos (*Bassaricyon*), with description of a new species, the Olinguito. *ZooKeys*, 324:1–83, 2013.

L. S. T. Ho and C. Ané. Asymptotic theory with hierarchical autocorrelation: Ornstein–Uhlenbeck tree models. *Ann. Stat.*, 41(2):957–981, 2013.

K. E. Jones, J. Bielby, M. Cardillo, S. A. Fritz, J. O'Dell, C. David L. Orme, K. Safi, W. Sechrest, E. H. Boakes, C. Carbone, C. Connolly, M. J. Cuttis, J. K. Foster, R. Grenyer, M. Habib, C. A. Plaster, S. A. Price, E. A. Rigby, J. Rist, A. Teacher, O. R. P. Binnida-Emonds, J. L. Gittleman, G. M. Mace, and A. Purvis. PanTHERIA: a species–level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology*, 90(9):2648, 2009.

A. Mooers, O. Gascuel, T. Stadler, H. Li, and M. Steel. Branch lengths on birth-death trees and the expected loss of phylogenetic diversity. *Syst. Biol.*, 61(2):195–203, 2012.

S. Sagitov and K. Bartoszek. Interspecies correlation for neutrally evolving traits. *J. Theor. Biol.*, 309:11–19, 2012.

A. Sofo. Harmonic number sums in higher powers. *J. Math. Anal.*, 2(2):15–22, 2011.

A. Sofo. New classes of harmonic number identities. *J. Int. Seq.*, 15:Art. 12.7.4, 2012.

A. Sofo. Finite number sums in higher order powers harmonic mumbers. *Bull. Math. Anal. Appl.*, 5(1):71–79, 2013.

T. Stadler. On incomplete sampling under birth-death models and connections to the sampling-based coalescent. *J. Theor. Biol.*, 261(1):58–68, 2009.

T. Stadler. Simulating trees with a fixed number of extant species. *Syst. Biol.*, 60(5):676–684, 2011.

T. Stadler and M. Steel. Distribution of branch lengths and phylogenetic diversity under homogeneous speciation models. *J. Theor. Biol.*, 297:33–40, 2012.

M. Steel and A. McKenzie. Properties of phylogenetic trees generated by Yule–type speciation models. *Math. Biosci.*, 170:91–112, 2001.

G. U. Yule. A mathematical theory of evolution: based on the conclusions of Dr. J. C. Willis. *Philos. T. Roy. Soc. B*, 213:21–87, 1924.

A Auxiliary results involving harmonic numbers

Some of the following results can be found in Adamchik [1997] and Sofo [2011, 2012, 2013].

Lemma A.1 *We have*

$$\sum_{k=1}^{n-1} \frac{H_k}{k(k+1)} = \bar{H}_n - \frac{H_n}{n}, \quad \sum_{k=1}^{\infty} \frac{H_k}{k(k+1)} = \frac{\pi^2}{6},$$

and for $m \geq 1$,

$$\sum_{k=1}^{n-1} \frac{H_k}{(k+m)(k+m+1)} = \frac{H_m}{m} - \frac{H_{n+m} - H_n}{m} - \frac{H_n}{n+m},$$
$$\sum_{k=1}^{n-1} \frac{\bar{H}_k}{(k+m)(k+m+1)} = \frac{n\bar{H}_n}{(n+m)m} - \frac{H_m}{m^2} + \frac{H_{n+m} - H_n}{m^2},$$

so that

$$\sum_{k=1}^{\infty} \frac{H_k}{(k+m)(k+m+1)} = \frac{H_m}{m}, \qquad \sum_{k=1}^{\infty} \frac{\bar{H}_k}{(k+m)(k+m+1)} = \frac{\pi^2}{6m} - \frac{H_m}{m^2}.$$

In particular,

$$\sum_{k=1}^{n-1} \frac{H_k}{(k+1)(k+2)} = \frac{n-H_n}{n+1}, \quad \sum_{k=1}^{n-1} \frac{H_k}{(k+2)(k+3)} = \frac{3n^2+5n-4(n+1)H_n}{4(n+1)(n+2)},$$

$$\sum_{k=1}^{n-1} \frac{\bar{H}_k}{(k+1)(k+2)} = \frac{n\bar{H}_n - n}{n+1}, \quad \sum_{k=1}^{n-1} \frac{\bar{H}_k}{(k+2)(k+3)} = \frac{n\bar{H}_n}{2(n+2)} - \frac{3n^2+5n}{8(n+1)(n+2)},$$

and

$$\sum_{k=1}^{\infty} \frac{H_k}{(k+1)(k+2)} = 1, \quad \sum_{k=1}^{\infty} \frac{H_k}{(k+2)(k+3)} = \frac{3}{4}, \quad \sum_{k=1}^{\infty} \frac{H_k}{(k+3)(k+4)} = \frac{11}{18},$$

$$\sum_{k=1}^{\infty} \frac{\bar{H}_k}{(k+1)(k+2)} = \frac{\pi^2}{6} - 1, \quad \sum_{k=1}^{\infty} \frac{\bar{H}_k}{(k+2)(k+3)} = \frac{\pi^2}{12} - \frac{3}{8}, \quad \sum_{k=1}^{\infty} \frac{\bar{H}_k}{(k+3)(k+4)} = \frac{\pi^2}{18} - \frac{11}{54}.$$

PROOF Clearly,

$$\sum_{k=1}^{n-1} \frac{H_k}{k(k+1)} = \sum_{k=1}^{n-1} \frac{1}{k(k+1)} \sum_{i=1}^{k} \frac{1}{i} = \sum_{i=1}^{n-1} \frac{1}{i}\left(\frac{1}{i} - \frac{1}{n}\right) = \bar{H}_n - \frac{H_n}{n}.$$

Similarly for $m \geq 1$, we have

$$\begin{aligned} \sum_{k=1}^{n-1} \frac{H_k}{(k+m)(k+m+1)} &= \sum_{i=1}^{n-1} \frac{1}{i}\left(\frac{1}{i+m} - \frac{1}{n+m}\right) = \sum_{i=1}^{n} \frac{1}{i}\left(\frac{1}{i+m} - \frac{1}{n+m}\right) \\ &= \frac{1}{m}\left(\sum_{i=1}^{n} \frac{1}{i} - \sum_{i=1}^{n} \frac{1}{i+m}\right) - \frac{H_n}{n+m} = \frac{1}{m}(H_n - H_{n+m} + H_m) - \frac{H_n}{n+m}, \end{aligned}$$

and

$$\begin{aligned} \sum_{k=1}^{n-1} \frac{\bar{H}_k}{(k+m)(k+m+1)} &= \sum_{i=1}^{n} \frac{1}{i^2}\left(\frac{1}{i+m} - \frac{1}{n+m}\right) = \frac{1}{m}\left(\sum_{i=1}^{n} \frac{1}{i^2} - \sum_{i=1}^{n} \frac{1}{i(i+m)}\right) - \frac{\bar{H}_n}{n+m} \\ &= \frac{1}{m}\left(\bar{H}_n - \frac{1}{m}(H_n + H_m - H_{n+m})\right) - \frac{\bar{H}_n}{n+m} = \frac{n\bar{H}_n}{(n+m)m} - \frac{H_m}{m^2} + \frac{H_{n+m} - H_n}{m^2}. \end{aligned}$$

□

Lemma A.2 *We have*

$$\sum_{k=1}^{n-1} \frac{H_k^2}{(k+1)(k+2)} = \bar{H}_n + \frac{n - H_n^2 - 2H_n}{n+1},$$

$$\sum_{k=1}^{n-1} \frac{H_k^2}{(k+2)(k+3)} = \frac{\bar{H}_n}{2} + \frac{11n^2+21n}{8(n+1)(n+2)} - \frac{H_n(2n+3)}{(n+1)(n+2)} - \frac{H_n^2}{n+2},$$

and generally for $m \geq 1$,

$$\sum_{k=1}^{\infty} \frac{H_k^2}{(k+m)(k+m+1)} = \frac{1}{m}\left(\frac{\pi^2}{6} + H_m^2 + \bar{H}_m - \frac{H_m}{m}\right).$$

In particular,

$$\sum_{k=1}^{\infty} \frac{H_k^2}{(k+1)(k+2)} = \frac{\pi^2}{6} + 1, \quad \sum_{k=1}^{\infty} \frac{H_k^2}{(k+2)(k+3)} = \frac{\pi^2}{12} + \frac{11}{8}, \quad \sum_{k=1}^{\infty} \frac{H_k^2}{(k+3)(k+4)} = \frac{\pi^2}{18} + \frac{37}{27}.$$

PROOF For $m \geq 1$,

$$\begin{aligned}
\sum_{k=1}^{n-1} \frac{H_k^2}{(k+m)(k+m+1)} &= \sum_{k=1}^{n-1} \frac{H_k}{(k+m)(k+m+1)} \sum_{i=1}^{k} \frac{1}{i} = \sum_{i=1}^{n-1} \frac{1}{i} \sum_{k=i}^{n-1} \frac{H_k}{(k+m)(k+m+1)} \\
&= \sum_{i=1}^{n-1} \frac{1}{i}\left(\frac{H_m}{m} - \frac{H_{n+m}-H_n}{m} - \frac{H_n}{n+m} - \frac{H_m}{m} + \frac{H_{i+m}-H_i}{m} + \frac{H_i}{i+m}\right) \\
&= \sum_{i=1}^{n-1} \frac{1}{i}\left(\frac{H_{i+m}-H_i}{m} + \frac{H_i}{i+m}\right) - \frac{H_{n-1}(H_{n+m}-H_n)}{m} - \frac{H_{n-1}H_n}{n+m} \\
&= \frac{1}{m} \sum_{i=1}^{n-1}\left(\frac{H_{i+m}}{i} - \frac{H_i}{i+m}\right) - \frac{H_{n-1}(H_{n+m}-H_n)}{m} - \frac{H_{n-1}H_n}{n+m}.
\end{aligned}$$

Observe that

$$\sum_{i=1}^{n-1}\left(\frac{H_{i+1}}{i} - \frac{H_i}{i+1}\right) = \sum_{i=1}^{n-1}\left(\frac{H_{i+1}}{i} - \frac{H_i}{i}\right) + \bar{H}_n - \frac{H_n}{n} = \bar{H}_n + 1 - \frac{1}{n} - \frac{H_n}{n},$$

and for $k \geq 2$,

$$\begin{aligned}
\sum_{i=1}^{n-1}\left(\frac{H_{i+k}}{i} - \frac{H_i}{i+k}\right) - \sum_{i=1}^{n-1}\left(\frac{H_{i+k-1}}{i} - \frac{H_i}{i+k-1}\right) &= \sum_{i=1}^{n-1} \frac{1}{i(i+k)} + \sum_{i=1}^{n-1} \frac{H_i}{(i+k)(i+k-1)} \\
&= \frac{1}{k}(H_{n-1} + H_k - H_{n+k-1}) + \frac{H_{k-1}}{k-1} - \frac{H_{n+k-1}-H_n}{k-1} - \frac{H_n}{n+k-1} \\
&= \frac{H_k}{k} + \frac{H_{k-1}}{k-1} - \frac{H_{n+k-1}-H_{n-1}}{k} - \frac{H_{n+k-1}-H_n}{k-1} - \frac{H_n}{n+k-1}.
\end{aligned}$$

It follows

$$\begin{aligned}
\sum_{i=1}^{n-1}\left(\frac{H_{i+m}}{i} - \frac{H_i}{i+m}\right) &= \bar{H}_n + 1 - \frac{1}{n} - \frac{H_n}{n} \\
&\quad + \sum_{k=2}^{m}\left(\frac{H_k}{k} + \frac{H_{k-1}}{k-1} - \frac{H_{n+k-1}-H_{n-1}}{k} - \frac{H_{n+k-1}-H_n}{k-1} - \frac{H_n}{n+k-1}\right) \\
&= \bar{H}_n - \frac{1}{n} - \frac{H_n}{n} + 2\sum_{k=1}^{m} \frac{H_k}{k} - \frac{H_m}{m} \\
&\quad - \sum_{k=2}^{m}\left(\frac{H_{n+k-1}-H_{n-1}}{k} + \frac{H_{n+k-1}-H_n}{k-1}\right) - H_n(H_{n+m-1}-H_n).
\end{aligned}$$

Using the classical relation $2\sum_{k=1}^{m}\frac{H_k}{k}=H_m^2+\bar{H}_m$ which follows from

$$\sum_{k=1}^{m}\frac{H_k}{k}=\sum_{k=1}^{m}\frac{1}{k}\sum_{i=1}^{k}\frac{1}{i}=\sum_{i=1}^{m}\frac{1}{i}\sum_{k=i}^{m}\frac{1}{k}=\sum_{i=1}^{m}\frac{H_m-H_{i-1}}{i}=H_m^2+\bar{H}_m-\sum_{k=1}^{m}\frac{H_k}{k},$$

we get

$$\begin{aligned}\sum_{i=1}^{n-1}\left(\frac{H_{i+m}}{i}-\frac{H_i}{i+m}\right)&=\bar{H}_n-\frac{1}{n}+H_m^2+\bar{H}_m-\frac{H_m}{m}\\&\quad-\sum_{k=2}^{m}\left(\frac{H_{n+k-1}-H_{n-1}}{k}+\frac{H_{n+k-1}-H_n}{k-1}\right)-H_n(H_{n+m-1}-H_{n-1})\\&=\bar{H}_n+H_m^2+\bar{H}_m-\frac{H_m}{m}-\sum_{k=1}^{m}\frac{H_{n+k-1}-H_{n-1}}{k}-\sum_{k=1}^{m-1}\frac{H_{n+k}-H_n}{k}-H_n(H_{n+m-1}-H_{n-1}),\end{aligned}$$

Thus

$$\begin{aligned}\sum_{k=1}^{n-1}\frac{H_k^2}{(k+m)(k+m+1)}&=\frac{1}{m}\left(\bar{H}_n+H_m^2+\bar{H}_m-\frac{H_m}{m}\right)-\frac{H_{n-1}(H_{n+m}-H_n)}{m}-\frac{H_{n-1}H_n}{n+m}\\&\quad-\frac{1}{m}\sum_{k=1}^{m}\frac{H_{n+k-1}-H_{n-1}}{k}-\frac{1}{m}\sum_{k=1}^{m-1}\frac{H_{n+k}-H_n}{k}-\frac{H_n(H_{n+m-1}-H_{n-1})}{m}.\end{aligned}$$

To finish the proof it remains to observe that $\frac{H_{n+k}-H_n}{k}\to 0$ as $n\to\infty$ for any fixed k.

□